klasifikasi

by Sumarno Sumarno

Submission date: 27-Sep-2021 11:53AM (UTC+0700)
Submission ID: 1658467382
File name: Classification_of_Student_Complaints.pdf (265.63K)
Word count: 2510
Character count: 14675

JOINCS (Journal of Informatics, Network, and Computer Science)
Vol 3 (2020)
Articles

Classification of Student Complaints with the Naive Bayes and Literature Methods

Klasifikasi Keluhan Mahasiswa dengan Metode Naive Bayes dan Sastrawi

Sumarno Departement of Informatics Universitas Muhammadiyah Sidoarjo
Mochamad Alfan Rosid Departement of Informatics Universitas Muhammadiyah Sidoarjo

E-Complaint Data is a collection of data which contains comments or complaints from students towards the University. Lots of comments - comments that occur in the university environment about the performance facilities of lecturers etc ... Text mining is also known as text data mining or knowledge search in a textual database is a semi-automatic process of extracting data patterns. The purpose of text mining is to get useful information from a collection of documents. In this study using the naïve Bayes method with TFIDF weighting features. The stages will be taken to determine its classification. First is taking data from the E-Complaint System then the data will go through the preprocessing stage using literary libraries, after going through the preprocessing stage the data will be divided into 2 namely training data and testing data. Then the training data will be carried out the TF-IDF weighting process up to the probability, if it has, the next is to process the testing data by determining priors. Next is the data testing stage between the testing data and the training data, the results of the testing data will come in the form of a predetermined category. The trial results show that the classification of complaints with the naïve Bayes algorithm and with the TF-IDF feature and literary libraries in the preprocessing process has an average accuracy that is quite high at 82%.

Pendahuluan

Text mining (menambang teks) merupakan analisis teks dimana sumber data biasanya di dapatkan dari dokumen , dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan, keterkaitan dan kelas antar dokumen. Metode yang akan digunakan dalam penelitian ini adalah metode *Naïve Bayes.* Manfat dari text mining adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen.

Algoritma *Naive Bayes* merupakan metode untuk menghitung nilai probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset [1] . Secara sederhana, pengelompokan *Naïve Bayes* mengan[illegible]ap adanya suatu fitur tertentu dalam sebuah kelas tidak terkait dengan adanya fitur lainnya. Keuntungan menggunakan Naïve Bayes adalah metode ini hanya membutuhkan dua data yaitu data training dan data Testing untuk menguji suatu data yang ingin diperoleh. Maka peneliti disini menggunakan metode ini untuk lebih meningkatkan nilai akurasi yang lebih akurat dari penelitian-penelitian terdahulu.

Pada penelitian yang dilakukan oleh Mochamad Alfan Rosid dkk, klasifikasi keluhan mahasiswa dilakukan dengan metode centroid based classifier dan dengan fitur TF-IDF-ICF. Di sistem informasi sebelumnya data diambil dari database aplikasi e-complaint Universitas Muhammadiyah Sidoarjo. Adapun hasil uji coba menunjukkan bahwa klasifikasi keluhan dengan algoritma centroid based classifier dan dengan fitur TF-IDF-ICF memiliki rata-rata akurasi yang cukup tinggi yaitu 79.5%[2].

TF-IDF merupakan menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan sebuah kata didalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi di dalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen. Selain pembobotan menggunakan TF-IDF peneliti juga menggunakan tahap *preprocessing* adapun tahap yang digunakan yaitu *stopwords removal* dan *stemming*[3].

Tahap yang digunakan ialah *stopwords removal* merupakan proses penghilangan kata tidak penting pada deskripsi melalui pengecekan kata-kata hasil parsing deskripsi apakah termasuk di dalam daftar kata tidak penting (stoplist) atau tidak. Yang kedua ialah tahap stemming merupakan tahap untuk mengubah token-token menjadi token yang berupa kata dasar. Tujuan dari tahap stemming adalah mengurangi jumlah[4].

Text mining juga dikenal dengan text data mining atau pencarian pengetahuan di basis data tekstual adalah proses yang semi otomatis melakukan ekstraksi dari pola data. Sumber data yang digunakan pada text mining adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Tujuan dari text mining adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen[5][6].

Metode Penelitian

Dalam melaksanakan penelitian ini, kami melakukan beberapa tahap, yaitu dimulai dari pengambilan data, tahap preprocessing, pembagian data, pembobotan, penerapan metode naïve bayes sampai data berhasil terklasifikasi. Adapun tahapan penelitian tersebut dapat digambarkan pada Figure 1 sebagai berikut:

Figure 1. *Tahapan Penelitian*

Dataset yang digunakan berjumlah 573 yang berasal dari data komentar/keluhan mahasiswa di aplikasi E-Complaint. Dataset ini meliputi katagori sebagai berikut :

1. Kemahasiswaan
2. DPAL (Direktorat Pengelolaan Aset dan Lingkungan)
3. DA (Direktorat Akademik)
4. DK (DIrektorat Keuangan)
5. DSTI (Direktorat Sistem dan Teknologi)
6. PERPUSTAKAAN
7. FAKULTAS
8. DRPM (Direktorat Riset dan Pengabdian Masyarakat)
9. LAIN-LAIN
10. BPM (Badan Penjaminan Mutu)

Contoh Dataset dapat dilihat pada Table 1 berikut:

Dokumen Keluhan	Kategori
UKM agar difasilitasi computer dan printer	Kemahasiswaan
Assalamualaikum Pengadaan inventaris alat musik tolong bantuaannya karna alat musik itu kebutuhan pokok buat UKM IKABAMA	Kemahasiswaan
Pak/bu seketariat UKM IKABAMA tidak ada studio beserta	Kemahasiswaan

sebagian alat musiknya.	
LCD/proyektor diperbaiki , gambarnya sudah tidak jelas di ruang 401 gedung C	DPAL
Pintu kelas 201 diperbaiki daun pintunya sudah putus jadi sulit buka pintunya	DPAL
Disediakan sound/speaker pada tiap tempat pelayanan agar terdengar kalau ada yang dipanggil	DPAL

Table 1. *Contoh data keluhan mahasiswa*

Tahap Prepocessing

Tahap Preprocessing disini adalah pembersihan data, fungsi dari tahap ini adalah agar akurasi yang didapat menjadi baik. Tahap Preprocessing terdiri dari 4 bagian[7][8], yaitu :

1. Tahap Case Folding : Tahap ini merupakan proses mengubah seluruh kalimat menjadi huruf kecil
2. Tahap Tokenizing : Tahap ini adalah sebuah proses penguraian deskripsi yang semula berupa kalimat menjadi sebuah kata
3. Tahap Stemming : Tahap ini merupakan proses mengubah kata menjadi kata dasarnya dengan menghilangkan imbuhan-imbuhan pada kata dalam dokumen atau mengubah kata kerja menjadi kata benda . Stem (akar kata) adalah kata inti setelah imbuhan dihilangkan (awalan dan akhiran)
4. Tahap Tagging :Tahap ini merupakan tahapan untuk mencari bentuk awal atau root dari tiap-tiap kata lampau atau kata dari hasil stemming

Pembobotan Kata

Pembobotan kata adalah sebuah proses dari pemberian nilai bobot berdasarkan term indeks. Ada beberapa pembobotan kata yang bisa digunakan, namun dipenelitian ini peneliti menggunakan pembobotan kata TF-IDF. TF IDF merupakan hasil perkalian dari Term Frequency atau jumlah kemunculan kata pada tiap dokumen serta Inverse Document Frequency atau kemunculan sebuah term dalam dokumen yang paling sedikit. Rumus dari TF IDF dapat dilihat pada Table 2 :

W □□ , □□ = W □□□□□□ , □□ □□ idf □□	(1)

Table 2.

$W_{,}$: Nilai *term frequency*

$idf_{□}$ Nilai *Inverse Document Frequency*

Term Frequency (TF)

Term frequency adalah seberapa sering kemunculan kata pada satu dokumen[9]. Rumus TF adalah pada Table 3 sebagai berikut :

W $tf\ i,d$ = { 1+log10 tf t,d , if tf t,d >0 0 }	(2)

Table 3.

Keterangan :

$Tf_{i,d}$ adalah jumlah kemuculan term t di dokumen d.

Document Frequency (DF)

Documenn requency merupakan kata yang sering muncul di dokumen[9]. Contoh : dan, di, atau, bisa

Invers Document Frequency (IDF)

Invers document frequency adalah kemungkinan munculnya term diseluruh dokumen[9]. Maka term yang jarang sekali muncul, nilai bobot IDF semakin besar. Berikut adalah rumus dari IDF pada Table 4:

idf t = log10+N/df (t)	(3)

Table 4.

Keterangan :

N : Jumlah dokumen teks

$Df_{(t)}$: Jumlah dokumen yang mengandung term t.

Term Frequency – Invers Document Frequency (TF-IDF)

Metode TF-IDF menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut[3]. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut.

Untuk rumus pembobotan text dengan metode TF-IDF pada: Table 5.

W $i_{,}j$ = tf $i_{,}j$. log(N/df i)	(4)

Table 5.

Dimana:

i: kata kunci

j: dokumen

tf : Banyaknya kata ditemukan dari i

dalam j

df : jumlah dokumen yang

mengandung i

N: total banyaknya dokumen

Naïve Bayes Classifier

Naïve Bayes Classifier merupakan metode yang berfungsi untuk mrnghitung nilai probabilitas

dengan menjumlahkan frekuensi dan kombinasi nilai dari data set[10]. Keuntungan menggunakan *Naïve Bayes* adalah metode ini hanya membutuhkan dua data yaitu data training (training set) dan data Testing (Testing set) untuk menguji suatu data yang ingin diperoleh. Rumus Naïve Bayes Classifier pada Table 6.

P (C*j* **V** W*i*) = P(C*j*) **x** P (W*i* **V** C*j*) **/** P(W*i*)	(5)

Table 6.

$P(C_j|W_i)$:*Posterior*, adalah kemunculan peluang pada kategori j tertentu ketika terdapat kemunculan kata i

$P(C_j)$:Prior, adalah peluang kemunculan dokumen pada kategori j

$P(W_i|C_j)$: *Likelihood* atau *Conditional Probability*, adalah peluang sebuah kata i masuk ke dalam kategori j

$P(W_i)$: Evidence, adalah peluang kemunculan sebuah kata

i : indeks kata yang berawal dari 1 sampai dengan kata ke-k

j : indeks kategori yang berawal dari 1 sampai dengan kategori ke-n

Menghitung jumlah dokumen pada kategori tertentu digambarkan pada persamaan dalam Table 7 berikut:

P(Cj) = N(Cj) / N	(6)

Table 7.

$N(C_j)$: jumlah dokumen latih yang masuk

dalam kategori j

N : jumlah keseluruhan dokumen

Multinomial Model merupakan model probabilitas yang penelti gunakan. Berikut merupakan persamaan *Multinomial Model* pada Table 8.

$P(w \text{ V } c) = \text{Count}(w,c)+1 \text{ / Count}(c)+ \lvert V\rvert$	(7)

Table 8.

Count(w,c) = jumlah kemunculan kata w pada kategori c

Count(c) = jumlah total kemunculan semua kata pada kategori c

|V| = jumlah term unik atau fitur

Metode Pengujian

Adapun tiga tahapan metode pengujian yang dipakai dalam penelitian ini adalah sebagai berikut :

Confusion matrix atau *error matrix* adalah sebuah metode perhitungan akurasi terhadap sebuah sistem pada konsep data mining. Terdapat 4 istilah di dalam *Confusion matrix* yaitu, True Positif (TP), True Negatif (TN), False Positif (FP), dan False Negatif (FN).

Precision merupakan tingkat ketepatan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem. Sedangkan *Recall* merupakan tingkat keberhasilan sistem menemukan kembali sebuah informasi.

1. Confusion Matrix
2. *Precision and Recall*
3. *Cross Validation*

Cross Validation adalah metode statistik untuk mengukur kinerja model algoritma dimana data dipisahkan menjadi 2 subset, yaitu data latih dan data uji. Pada penelitian ini penulis menggunakan 10 k-fold cross validation.

Hasil Penelitian dan Pembahasan

Hasil dari penelitian ini adalah sebuah sistem klasifikasi keluhan mahasiswa yang dapat secara otomatis mengklasifikasi keluhan-keluhan mahasiswa ke unit kerja tujuan keluhan. Adapun tampilan dari sistem yang dihasilkan dapat dilihat pada Figure 2 berikut:

Figure 2. *Tampilan data keluhan mahasiswa*

Sebelum sistem ini dapat mengklasifikasi data keluhan mahasiswa, harus melalui tahap *preprocessing* terlebih dahulu melalui menu *preprocessing* yang ditunjukkan oleh Figure 3.

Figure 3. *Tampilan Halaman Preprocessing*

Setelah melalui tahap proses *preprocessing* data dibagi menjadi 2 yaitu data *training* dan data *testing* seperti Figure 4 berikut:

Figure 4. *Pembagian data training dan testing*

Pada pembagian data ini, data keluhan akan diambil secara acak. Pada penelitian ini menggunakan data training 70% dari *dataset*. Kemudian akan dicoba klasifikasi data testing terhadap data training yang sudah melalui pembobotan dan perhitungan probabilitas seperti Figure 5 berikut:

Figure 5. *Hasil klasifikasi data testing*

Dari Figure 5 terlihat bahwa sistem berhasil melakukan klasifikasi keluhan mahasiswa, ada yang dapat terklasifikasi dengan benar sesuai dengan klasifikasi manual, ada yang terklasifikasi salah. Untuk mengukur nilai akurasi dilakukan dengan menggunakan metode cross validation, recall and precission dan confusion matrix, pada penelitian ini hasil ketiga metode tersebut dapat dilihat pada Figure 6, Figure 7, Figure 8 sebagai berikut:

Figure 6. *Tampilan uji coba dengan metode confusion matrix*

Figure 7. *Tampilan uji coba dengan precission dan recall*

Figure 8. *Tampilan uji coba dengan metode cross validation*

Terlihat bahwa pada hasil uji coba, sistem yang dihasilkan memiliki tingkat akurasi rata-rata 82%.

Kesimpulan

Metode *naïve bayes* dan library sastrawi dapat melakukan klasifikasi keluhan mahasiswa dengan baik dan dapat memperbaiki nilai akurasi pada penelitian sebelumnya yang memiliki nilai akurasi rata-rata 79% menjadi 82% menggunakan *naïve bayes* dan library sastrawi pada proses *preprocessing*-nya.

References

1. A. S. Fitriani, T. Informatika, F. Teknik, and U. M. Sidoarjo, “Penerapan Data Mining Menggunakan Metode Klasifikasi Naïve Bayes untuk Memprediksi Partisipasi Pemilihan Gubernur,” vol. 3, no. 2, pp. 98–104, 2019.
2. M. A. Rosid, G. Gunawan, and E. Pramana, “Centroid Based Classifier With TF – IDF – ICF for Classfication of Student’s Complaint at Appliation E-Complaint in Muhammadiyah University of Sidoarjo,” vol. 1, no. 1, 2015.
3. R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, “Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi,” J. Tek. Elektro, vol. 9, no. 1, pp. 18–23, 2017.
4. A. Rachmat C and Y. Lukito, “Klasifikasi Sentimen Komentar Politik dari Facebook Page Menggunakan Naive Bayes,” J. Inform. dan Sist. Inf. Univ. Ciputra, vol. 2, no. 2, pp. 26–34, 2016, doi: 10.1080/10408398.2013.809690.
5. S. Gusriani, K. D. K. Wardhani, and M. I. Zul, “Analisis Sentimen Terhadap Toko Online di Sosial Media Menggunakan Metode Klasifikasi Naïve Bayes (Studi Kasus: Facebook Page BerryBenka),” 4th Appl. Bus. Eng. Conf., vol. 1, no. 1, pp. 1–7, 2016.
6. Spr. Rani, B. Ramesh, and M. Anusha, “Evaluation of stemming techniques for text classification,” J. Comput. …, vol. 43, no. 3, pp. 165–171, 2015.
7. J. T. Informasi et al., “PENGARUH TEXT PREPROCESSING DAN KOMBINASINYA,” vol. 15, pp. 1–11, 2019.
8. S. Vijayarani, M. J. Ilamathi, M. Nithya, A. Professor, and M. P. Research Scholar, “Preprocessing Techniques for Text Mining -An Overview,” vol. 5, no. 1, pp. 7–16.
9. W. E. Nurjanah, R. S. Perdana, and M. A. Fauzi, “Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan M,” no. October, 2017.
10. S. S. Nikam, “A Comparative Study of Classification Techniques in Data Mining Algorithms,” Int. J. Mod. Trends Eng. Res., vol. 4, no. 7, pp. 58–63, 2017, doi: 10.21884/ijmter.2017.4211.vxayk.

klasifikasi

ORIGINALITY REPORT

19%	19%	7%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Muhammadiyah Sidoarjo Student Paper	6%
2	id.123dok.com Internet Source	4%
3	journal.ummat.ac.id Internet Source	3%
4	id.scribd.com Internet Source	1%
5	www.scribd.com Internet Source	1%
6	ejournal.bsi.ac.id Internet Source	1%
7	etheses.uin-malang.ac.id Internet Source	1%

Exclude quotes	On	Exclude matches	< 30 words
Exclude bibliography	On		